ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/380011781

Limits of Automata—Then and Now: Challenges of Architecture, Brittleness, and Scale

Article *in* Journal of Cognitive Engineering and Decision Making · April 2024

DOI: 10.1177/15553434241240203

CITATIONS
5

READS
1,856

1 author:

David D Woods
The Ohio State University
447 PUBLICATIONS **30,124** CITATIONS

SEE PROFILE

All content following this page was uploaded by David D Woods on 23 April 2024.

The user has requested enhancement of the downloaded file.

Check for updates

Special Issue Paper

Limits of Automata—Then and Now: Challenges of Architecture, Brittleness, and Scale

Journal of Cognitive Engineering and Decision Making
2024, Vol. 0(0) 1–8
© 2024, Human Factors and Ergonomics Society
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/15553434241240203
journals.sagepub.com/home/edm
Sage

David D. Woods[1]

Abstract

Two trajectories underway transform human systems. Processes of growth/complexification have accelerated as stakeholders seek advantage from advances in connectivity/autonomy/sensing. Surprising empirical patterns also arise—puzzling collapses of critical valued services occur against a background of growth. In parallel, new scientific foundations have arisen from diverse directions explaining the observed anomalies and breakdowns, highlighting basic weaknesses of automata regardless of technology. Conceptual growth provides laws, theorems, and comprehensive theories that encompass the interplay of autonomy/people and complexity/adaptation across scales. One danger for synchronizing the trajectories is *conceptual lag* as researchers remain stuck in stale frames unable to keep pace with transformative change. Any approach that does not either build on the new conceptual advances—or provide alternative foundations—is no longer credible to match the scale and stakes of modern distributed layered systems and overcome the limits of automata. The paper examines longstanding challenges by contrasting progress *then* as the trajectories gathered steam, to situation *now* as change has accelerated.

Keywords

autonomy, scale, complexity < topics, adaptive capacity, joint cognitive system, brittleness, limits of autonomy

Introduction

"Not Even Wrong"

Pieces like the target article remind me of a story about the 20th century physicist Wolfgang Pauli (Peierls, 1992). A colleague had received a draft paper and pressed Pauli to examine it. Reluctantly, Pauli began to look through the paper. Then he handed it back saying simply, "not even wrong."

As fields develop, conceptual lag sets in, even as knowledge grows in the form of empirical generalizations, laws, and formal theories (Dunbar, 1993). The knowledge advances bring new light to challenges and provide new paths forward in understanding fundamentals. Pauli's remark cuts to the heart of the lag problem: some work in science and engineering simply falls behind the pace of challenge and response (Woods & Alderson, 2021).

The dangers from conceptual lag apply to challenges at the intersection of people, technology, and work as exemplified by the target article.

[1]Institute for Ergonomics with Cognitive Systems Engineering Laboratory, OH, USA

Corresponding Author:
David D. Woods, Institute for Ergonomics with Cognitive Systems Engineering Laboratory, 210 Baker Syst., 1971 Neil Ave, Columbus, OH 43210-1271, USA.
Email: woods.2@osu.edu

As new technology is developed and deployed, processes of growth and complexification are stimulated. New challenges arise; new versions of old challenges come to the fore; old challenges take on new urgency because systems expand to new scales with extensive interdependencies across layers and roles.

Connected to older foundations and derived from diverse lines of inquiry, a variety of conceptual developments have been going on to match the new challenges (see supplemental material):

(a) laws generalized from empirical patterns observed across multiple settings (e.g., Fluency Law; Law of Stretched Systems; Woods & Hollnagel, 2006),
(b) proven and provable theorems e.g., the Robust yet Fragile Theorem (Carlson & Doyle, 2000), and
(c) comprehensive formal theories that encompass autonomy, people, complexity, and adaptation (the two currently in play are Diversity Enabled Sweet Spots, (Nakahira et al., 2021), and the Theory of Graceful Extensibility, (Woods, 2018)

These twin developments—complexity penalties accompanying growth and progress on scientific foundations of adaptive systems at human/technological scales—create the conditions for conceptual lag. Pieces like the target article rehashing ad hoc taxonomies/guidelines and laboratory replications of well-established real-world findings may have had value in the distant past. Now they lag the trajectory and scope of conceptual advances. Critically, they lag the trajectory and scale of challenges that accompany growth—growth stimulated by technological advances, including automata, that has profoundly transformed the nature of systems (Woods & Alderson, 2021).

Four challenges arise from the trajectory of growth.

- The *problem of architecture* has engulfed previous concerns on the design of human-automation/autonomy interaction (Johnson et al., 2014; Supplemental material on Progress on the Architecture Problem).
- The *risk of brittle collapse* is universal and out-of-control in the deployment of autonomous capabilities (Woods, 2018).
- Operations go on at *multiple scales in parallel*, especially in the face of disrupting events. The rise of highly autonomous critical digital services which provision the delivery of valued services to diverse stakeholders in all sectors is the exemplar (Supplemental material on Trajectories of Complexification).
- Throughout all of the changes, the limits and *weaknesses of algorithms*/machines continue to be overlooked, beginning with Norbert Wiener's (1950) warnings (Wiener, 1950); ostensibly the point of the target article (Supplemental material on Risks of Automata Alone: General & Sample Cases).

The new scale of systems, fueled partially by increasingly autonomous capabilities, has super charged the dangers of literal-minded machines. The result is the drive for new architectures to synchronize layered distributed networks of many interacting and interdependent human and machine roles. The conceptual advances provide the roadmap to architect such layered distributed systems. As a result, the science underlying systems engineering has undergone a sea change. Any approach that does not build on these advances—or provide alternative competing foundations—is not credible given the scale and stakes of modern distributed layered systems.

The Architecture Problem

Then

Hollnagel and Woods (1983)/Woods (1985) introduced the architecture problem for a previous cycle of change driven by Artificial Intelligence (AI) and because large failure events led to increased pressure to deploy automation as a response to accidents throughout the 1980s. The key points, also raised by others (e.g., Hutchins, 1995), include the following:

(1) Introducing new technology is based on, or changes, a *joint system* architecture;

(2) the common architecture is implicit for the most part and built on the assumption of substitution of machines for people;
(3) this common implicit architecture is poor on several grounds; among these was the difficulties the architecture imposes when the automation/AI was *wrong or when events fell outside of its competence envelope* (Roth et al., 1987; Supplemental material on Risks of Autonomy Alone);
(4) paths forward concern how to design joint systems that carry out difficult forms of cognitive and distributed work—joint activity (Supplemental material on Progress on the Architecture Problem).

Empirical corroboration followed as projects were commissioned to repair systems when the weaknesses became apparent following deployment. This led to work on principles and techniques for alternative joint architectures. One technical anchor was modeling the demands on cognitive work that any set of human and machine roles would need to handle (The Law of Demands; Woods & Hollnagel, 2006, p. 171). A second anchor was how machines could participate in joint activity (Klein et al., 2005). How can an autonomous machine act as a cooperative and adaptive partner when new events disrupt ongoing organized activities?

Early efforts demonstrated that one could design joint systems that performed better than the common implicit architecture (see Supplemental material on Risks of Automata: Early Cases). One early example began because automation alone performed poorly (failures with high economic costs) on a difficult real-world case (the fundamental difficulty was not widely recognizes). A few people developed the skill to carry out the critical operation, though setbacks still occurred. The latest advance in algorithms—symbolic AI—seemed to provide the opportunity to build new automation that would overcome past weaknesses in machine performance and eliminate the need for human expertise.

Instead the researchers started with the Law of Demands—diagnose what makes this form of work hard for any joint cognitive system (Roth & Woods, 1988). This approach revealed the work required dynamic synchronization across three roles, each responsible for a portion of three highly interdependent nonlinear processes. The complex dynamics, plus limits on sensing, meant operators, human, and machine, were operating "blind" about the effects of actions across the three interdependent processes. Simply replacing the older algorithm/automation with new algorithms could not address the difficulties. Anticipatory information was essential; a novel mathematical twist generated the missing information. Context-sensitive advice on how to outmaneuver the complex dynamics was also needed to support coordination across the three roles in the joint system (Woods & Roth, 1988).

Now

More work followed by various groups expanding the techniques and demonstrating the value of a joint system approach (see Supplemental material on Co-Active Architecture). However, the common implicit architecture still continues to dominate deployments of new autonomous capabilities. What went wrong?

The line of work showed one *could* build different joint architectures. The impact was limited by the inability to produce *computational and scalable* alternative architectures that could become part of system engineering organizations given resource pressures.

Building a joint architecture for a particular case of a sub-class of problems looked expensive. Each project had limited generality to accelerate progress on the next sub-class of problems when compared to the allure of the standard architecture. All assumed each new technology was a step toward the eventual realization of full automation with minimal side effects and only residual limits, even as the empirical findings grew that automation's limits mattered and an explicit joint approach was valuable.

Another impediment was the lack of infrastructure and tooling to make it practicable for non-specialists to build a reusable joint architecture. Computational tooling is required to scale cognitive work systems. This turned out to be critical because technological advances stimulated growth leading to broader systems operating through many interacting automata with extensive interdependencies. Most deployed systems consist of multiple human roles at different system layers

all assisted by a heterogenous mix of technologies and automation.

The Risk of Brittleness

Then

The joint system perspective quickly led to studies that highlighted the brittleness of automata regardless of the specific technology used. What is brittleness? Descriptively, brittleness is how rapidly a system's performance declines when it nears and reaches its boundary. Brittle systems experience rapid performance collapses, or failures, when events challenge system boundaries.

Roth et al. (1987) examined a diagnostic AI system built (and ready to deploy) according to the standard development process for the common implicit joint architecture. Actual problems were injected into a working example of the system which had been removed from service. Actual technicians were supposed to follow guidance from the AI system to carry out observations, tests, and enter data. Presumably, the AI system would diagnose problems leaving technicians to execute AI-directed repair actions.

What actually happened is the diagnostic process controlled by the algorithm kept breaking down requiring significant engagement of the human technician, even though the faults were not extremely difficult. In reality, a variety of complicating factors arose which went outside the *competence envelope* of the AI system. Furthermore, everyone involved in the development process underestimated the occurrence of complicating factors and overestimated the algorithm's competence envelope. The system was brittle at the boundaries. People needed to be active in the diagnostic process. Active technicians invented ways to direct the AI algorithm as a tool even though the system deployed deliberately contained no mechanisms for directability. As the responsible party on the scene of the broken device/lost service, technicians adapted to fill the gaps in the developer's work. These findings were repeated with different forms of technology for different forms of cognitive work in different sectors, but all with the same basic architecture.

These studies led to ways to design better joint systems to mitigate brittleness and other weaknesses of automata. One notable successful program to develop an alternative and effective joint system architecture has been led by Matt Johnson (Johnson et al., 2014) including building/deploying a joint system design to meet the demanding performance criteria of one of the DARPA Grand Challenge programs (Supplementary materials on Co-Active Architecture).

Now

The risk of brittle collapse is evident all around us in today's worlds which operate at scale with extensive interdependencies across layers, roles, and organizations. The list of events is extensive and expands regularly. Notable examples of brittleness come in the form of *runaway automation* cases such as the twin Boeing 737 Max accidents in 2018/2019 (three additional fatal aviation accidents involved automation "misbehavior" based on sensor failures) or to the 2012 collapse of the Knight Capital financial institution from runaway automated trading producing large financial losses (see Supplemental materials on Risks of Automata: Cases). In these failures, interactions across multiple system layers were significant in the genesis and inability to resolve these breakdowns.

The new work revealed how automata are competent and brittle and highlighted how joint systems were competent and extensible. Critically, competence and brittleness are separate properties. In 2000, Carlson and Doyle explain the empirical patterns by discovering and proving the *Robust Yet Fragile theorem*: systems "are robust to perturbations they were *designed to handle*, yet fragile to unexpected perturbations and design flaws" (p. 2529). Empirically, studies of cognitive work systems subject to large failures also discovered brittleness is fundamental: modern systems are highly competent when events fall within the envelope of designed-for-uncertainties, but subject to sudden, large failures occur in the face of events that challenge or go beyond the envelope. The theoretical progress on distributed layered systems has revealed how design inevitably will leave weaknesses in autonomous sub-systems and how perturbations beyond the competence envelope are guaranteed (Supplemental material Progress on Architecting Distributed Layered Networks of People and Automata).

The combined work revealed the risk of brittle collapse is *universal* for all systems at all scales, including systems which depend critically on deployed autonomous subsystems. Today the risk of brittle collapse is *out-of-control* in systems that depend on deployed autonomous subsystems.

Essentially, a difference in degree has led to a difference in kind—we have been deploying partial but increasingly autonomous systems with partial but increasing authority for decades until these systems have morphed into *distributed layered networks* with extensive, hidden, and tangled interdependencies.

Growth, Complexification, and Scale

Now

One world instantiates the dramatic scale shift that has occurred through processes of growth and complexification stimulated by the widespread deployment of automata—critical digital services (Woods, 2024). University–industry partnerships have enabled studies of the flow of incidents/outages and how incidents are usually handled well in this sector—processes which underpin every sector today (Woods, 2017). The findings have revealed patterns about how layered networks of automata at scale function/malfunction and on the forms of cognitive work that occur in systems operating at this scale (Suplementary material on Trajectories of Complexification).

Critical digital services work at scale through partially autonomous software functions which act with relatively high authority and interact with other automata. The network of automata encompasses a layered network of human roles, human organizations, human stakeholders responsible for critical infrastructures, and the services supported. Consider the multiple airline service outages (including outages at British Airways, Delta, Southwest, and American) that undermined the ability to fly passengers for extended blocks of time costing hundreds of millions of dollars plus the disrupting effects on travelers' goals. These are all breakdowns in highly automated, highly autonomous, layered systems at scale.

Growth fueled by widespread deployments of increasingly autonomous subsystems changes the landscape of how trouble arises and spreads in systems that provide valued services usually reliably. New patterns of breakdown that threaten loss of control and outages emerge from the limits of automated subsystems and from weaknesses arising from the extensive interdependencies across subsystems. One is the *risk of saturation* where an automated process runs out of the ability to respond to increasing demands. As a process approaches saturation it can behave in ways that spread overload to other units expanding the risk of saturation across the network and undermining the ability of the whole to cope with the disturbances underway. Another is the oversubscription of a critical resource as different units respond to the effects of spreading disruptions (one variant occurs during recovery where many processes try to come back into action simultaneously producing overloads that exacerbate service outages).

"I didn't know it worked that way" is the most common statement during incident response/after-action reviews. The comment doesn't refer to a novel specific failure. Rather the comment says they didn't understand how the distributed system *actually worked* given interdependencies across parts and layers. In many of the incidents studied, there was no failure per se. Rather a change initiated for good reasons sets up interactions that came into play destructively—a breakdown pattern that emerges from the complexity penalties that accompany growth. Even when specific failures did occur, the system design was thought to have been robust to such failures. The design for robustness turned out to be limited in the face of the emergent complexities just as the general theories explain. The emergent (real) properties of layered networks mean breakdowns cannot be understood in terms of component or subsystem performance—"failure" and "success" at the new scale are entirely different.

The processes of growth and complexification mean the essential case for human-autonomy is no longer a human role using an automated system—the joint system now is layered, distributed, and tangled. The scale growth creates hidden interdependencies and effects at a distance that challenge diagnosis. Coordination costs rise as resolving incidents spreads over multiple

organizations. Anomalies produce high uncertainty as disturbances propagate. Activities spread across multiple human roles and organizations engaged in providing valued services. All human roles are supported by multiple forms of technology—heterogenous over capabilities, technologies, and robustness. Human supervision is spread over multiple roles concerned with a configuration of automated activities when disturbances arise. Deploying new technology impacts multiple roles, creates new roles, and expands relationships across problem holders and stakeholders at different system layers and contexts.

Automata are Limited (always)

Then

The studies of joint system quickly coalesced onto the theme of how people adapt to cope with complexity. Systems inevitably are messy with gaps; people provide the adaptive capacity to bridge the gaps. Designing joint systems provides support for people adapting to different forms of complexity. Empirical patterns were captured that cut across varying contexts and technology. General laws about macro-cognitive work systems emerged that applied broadly collated originally in (Woods & Hollnagel, 2006).

Now

Syntheses across multiple, usually disconnected, lines of inquiry have led to major conceptual advances to respond to the challenges of our turbulent world. Proven theorems and comprehensive formal theories have emerged to explain phenomenon, patterns, and empirical laws. Two comprehensive theories now provide a scientific foundation to architect-distributed layered human/technological systems (Supplemental material on Progress on Architecting Distributed Layered Networksof People and Automata).

The theories tackle the question: What architectures can cope with limits on automata (actually on all adaptive units)? Even if designed as well as possible given the constraints on development, the integrated system will be composed of multiple automata whose performance will represent different positions in a multi-dimensional trade space—accurate versus fast; thorough versus efficient; optimal now versus extensible in the future. This and other inescapable limits create the new challenge: how to architect a *layered* system of many partially autonomous roles of people and machines pursuing a variety of partially incompatible goals, so that the whole performs flexibly and adaptively over a wider set of contexts, disturbances, and goal conflicts as compared to subunits operating with specialized capabilities/and finite resources—always valuable but bounded. The theories specify how to architect networks of adaptive systems that can outmaneuver complexity.

The new conceptual advances are pragmatic. First, the new fundamentals define poor architectures—exposing the inherent weakness of the common implicit architecture that still dominates development/deployment of autonomous capabilities. Second, they provide architecting solutions that can meet the scale of today's problems. This involves automata which work and interact with each other and with human roles in completely different ways than today—how to behave when approaching saturation, how to demonstrate reciprocity when neighbors are under increasing stress, how to sacrifice and reprioritize in the face of goals conflicts, and how to reconfigure and synchronize activities as disruptions occur. Third, the architectural advances change the nature of system development, deployment, and operations. The model demonstrated practically but incompletely in the growth critical digital services turns out to be what the theories say is necessary—fundamentals that enable architecting for continuous adaptability. Fourth, new kinds of mathematical and computational tooling emerge that can escape the limits of custom design work that has plagued joint system design (Farjadian et al., 2021).

This is not the place to review all of the recent advances, nor is this the place to cover all of the complexity penalties that arise from growth fueled by new autonomous capabilities. The issue here is the dangers from conceptual lag. The growth and complexification of systems due to connectivity, autonomy, and sensing have increased the economic and safety consequences of failures as the resulting services provided become more valuable

and essential to a wider range of stakeholders. The stakes have been raised: interdependencies/scale have grown with technology advances; the costs of poor architectures and of competent but brittle automata have risen in parallel. To match the stakes, we now have proven theorems/formal theories that aspire to comprehensive scope—substantial scientific foundations. From these, we have practical alternative architectures, as well as practical demonstrations that work under real complexity.

The process of conceptual change is well underway. Any approach that does not build on/extend these advances, or provide alternative foundations, is not credible. Approaches that remain stuck in variations on the common implicit architecture will repeat the empirical patterns of puzzling brittle failures against a background of growth. Society's needs beckon.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

David D. Woods https://orcid.org/0000-0003-3131-2894

Supplemental Material

Supplemental material for this article is available online.

References

Carlson, J. M., & Doyle, J. C. (2000). Highly optimized tolerance: Robustness and design in complex systems. *Physics Review Letters*, *84*, 2529–2532. https://doi.org/10.1103/PhysRevLett.84.2529

Dunbar, K. (1993). Concept discovery in a scientific domain. *Cognitive Science*, *17*(3), 397–434. https://doi.org/10.1207/s15516709cog1703_3

Farjadian, A. B., Thomson, B., Annaswarmy, A. M., & Woods, D. D. (2021). Resilient flight cotnrol: An architecture for human supervision of automation. *IEEE Transactions on Control Systems Technology*, *29*(1), 29–42. https://doi.org/10.1109/TCST.2019.2959542

Hollnagel, E., & Woods, D. D. (1983). Cognitive systems engineering: New wine in new bottles. *International Journal of Man-Machine Studies*, *18*(6), 583–600. https://doi.org/10.1016/S0020-7373(83)80034-0

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, *3*(1), 43–69. https://doi.org/10.5898/JHRI.3.1.Johnson

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in Joint activity. In *Organizational Simulation* (pp. 139–184). John Wiley & Sons.

Nakahira, Y., Liu, Q., Sejnowski, T. J., & Doyle, J. C. (2021). Diversity-enabled sweet spots in layerred architectures and sped-accuracy trade-offs in sensorimotor control. *Proceedings of the National Academy of Science PNAS*, *118*(22). https://doi.org/10.1073/pnas.1916367118

Peierls, R. (1992). Where Pauli made his 'wrong' remark. *Physics Today*, *45*(12), 112. https://doi.org/10.1063/1.2809934

Roth, E. M., Bennett, K. B., & Woods, D. D. (1987). Human Interaction with and "intelligent" machine. *International Journal of Man-Machine Studies*, *27*(5–6), 479–525. https://doi.org/10.1016/S0020-7373(87)80012-3

Roth, E. M., & Woods, D. D. (1988). Aiding human performance: I. Cognitive analysis. *Le Travail Humain*, *51*(1), 39–64. https://www.jstor.org/stable/40657481

Wiener, N. (1950). *the human use of human beings: Cybernetics and society.* Doubleday.

Woods, D. D. (1985). Cognitive Technologies: The design of joint human-machine cognitive systems. *AI Magazine*, *6*(4), 86–92. https://doi.org/10.1609/aimag.v6i4.511

Woods, D. D. Ed. (2017). *STELLA report from SNAFU catchers workshop on coping with complexity* (*March*). SNAFU Catchers Consortium. https://stella.report

Woods, D. D. (2018). The theory of graceful extensibility. *Environment Systems and Decisions*, *38*, 433–457. https://doi.org/10.1007/s10669-018-9708-3

Woods, D. D. (2024). The science and pragmatics of resilience engineering through the lens of complexification. In *Series on Foundations of Resilience Enginnering* (*March*). Cognitive Systems Engineering Laboratory. https://www.youtube.com/watch?v=R0BEQETj9D0&t=36s

Woods, D. D., & Alderson, D. A. (2021). Progress toward Resilient Infrastructures: Are we falling behind the pace of events and changing threats? *Journal of Critical Infrastructure Policy*, *2*(2), 5–18. https://doi.org/10.18278/jcip.2.2.2

Woods, D. D., & Hollnagel, E. (2006). *Joint cognitive systems: Patterns in cognitive systems engineering*. Taylor & Francis.

Woods, D. D., & Roth, E. M. (1988). Aiding human performance: II. From cognitive analysis to support systems. *Le Travail Humain*, *51*(2), 139–171. https://www.jstor.org/stable/40657488

David Woods is Professor Emeritus in Integrated Systems Engineering at the Ohio State University and Past-President of the Resilience Engineering Association and Human Factors and Ergonomics Society. He is a pioneer in Resilience Engineering—how to support people as they adapt to cope with complexity. His books include Behind Human Error, Resilience Engineering, and Joint Cognitive Systems. He started the SNAFU Catchers industry-university consortium to apply the new science to build resilience in critical digital services.

Supplementary Material for Woods, "Limits of Automata Then and Now: Challenges of Architecture, Brittleness, and Scale."

General

Wiener, N. (1950). *The human use of human beings: Cybernetics and society*. NewYork: Doubleday.

Wiener, E. L. (1989). Human factors of advanced technology ("glass cockpit") transport aircraft. (NASA Contractor Report No. 177528). Moffett Field, CA: NASA-Ames Research Center.

Woods, D. D. (1989). The effects of automation on the human's role. In S. Norman and H. Orlady (Eds.), *Flightdeck Automation. Promises and Realities.* NASA. Ames Research Center, Moffett Field, CA, 1989, pp. 61-85.

Woods, D. D. (1996). Decomposing Automation: Apparent Simplicity, Real Complexity, In R. Parasuraman and M. Mouloula, editors, *Automation Technology and Human Performance: Theory and Applications,* Erlbaum, p. 3-17.

Murphy, R. R. and Woods, D. D. (2009). Beyond Asimov: The Three Laws of Responsible Robotics. *IEEE Intelligent Systems,* 24(4), July/August, pp. 14-20. [Reprinted in W. Wallach and P. M. Asaro, *Machine Ethics and Robot Ethics*. Routledge, 2017, 1st edition, 405-413.]

Murphy, R. & Shields, J. (2012). *The Role of Autonomy in DoD Systems*. Defense Science Board, Department of Defense, July. Available online: https://fas.org/irp/agency/dod/dsb/autonomy.pdf. (accessed on 5 February 2016).

Bradshaw, J., Feltovich, P., Hoffman, R. R., Johnson, M., and Woods, D. D. (2013). The Seven Deadly Myths of "Autonomous Systems". *IEEE Intelligent Systems*, 28(3), May/June, 54-61.

Woods, D. D. (2016). The Risks of Autonomy: Doyle's Catch. *Journal of Cognitive Engineering and Decision Making*, 10(2), 131-133.

Smith, P.J. (2018). Making Brittle technologies Useful. In P. Smith and R. Hoffman (Eds.) *Cognitive Systems Engineering: A Future for a Changing World*. Taylor and Francis, 181-208.

Johnson, M. & Vera, A. (2019). No AI Is an Island: The Case for Teaming Intelligence. *AI Magazine*, 40, (1/Spring).

Schraagen, J.-M. and van Diggelen, J. (2021). A Brief History of the Relationship between Expertise and Artificial Intelligence. In M.-L. Germain & R.S. Grenier, Eds., *Expertise at work: Current and emerging trends*. Cham: Palgrave Macmillan.

Woods, D.D. (2024). The Command-Adapt Paradox: : How Organizations are Slow and Stale to Adapt in a Turbulent World. In Le Coze, J.-C. and Journé, B. (Eds.), *Compliance and Initiative in the Production of Safety: A Systems Perspective on Managing Tensions and Building Complementarity.* Springer Open & the Foundation for Industrial Safety Culture, Springer Briefs in Safety Management, Toulouse France.

Sample of Cases

Roth, E. M., Bennett, K. and Woods, D.D. (1987). Human interaction with an 'intelligent 'machine. *International Journal of Man-Machine Studies*, 27:479–525.

Cook R. I., Woods, D. D. and Howie, M. B. (1992). Unintentional delivery of vasoactive drugs with an electromechanical infusion device. *Journal of Cardiothoracic and Vascular Anesthesia*, 6:238–244.

Leveson N. G. and Turner, C. S. (1993). An investigation of the Therac-25 accidents. *Computer*, 26(7), 18-41. 10.1109/MC.1993.274940

Sarter, N., Woods, D. D. and Billings C. E. (1997). Automation Surprises. In G. Salvendy, editor, *Handbook of Human Factors/Ergonomics*, second edition, Wiley, New York, pp. 1926-1943, 1997. (Reprinted in N. Moray, editor, Ergonomics: Major Writings. Taylor & Francis, 2004.

Smith, P.J., McCoy, E. and Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *IEEE Transactions on Systems, Man and Cybernetics*, 27, 360-371.

https://www.researchgate.net/publication/3411796_Brittleness_in_the_design_of_cooperative_problem-solving_systems_The_effects_on_user_performance

Knight Capital Runway Automation in Layered System

Allspaw, J. (2019). Hindsight and Sacrifice Decisions. Blog post March 3, 2019 https://www.adaptivecapacitylabs.com/2019/03/03/hindsight-and-sacrifice-decisions/

Hochstein, L. (2014). Counterfactual Thinking, Rules, and The Knight Capital Accident. Blog Post October 29, 2014 https://www.kitchensoap.com/2013/10/29/counterfactuals-knight-capital/

Progress on the Architecture Problem

Early Examples of Designing for Effective Joint Systems

Sorkin, R. D. and Woods, D.D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction*, 1:49--75.

Roth E.M. and Woods D.D. (1988). Aiding human performance: I. Cognitive analysis. *Le Travail Humain,* 51(1):39--64.

Woods D.D. and Roth. E.M. (1988). Aiding human performance: II. From cognitive analysis to support systems. *Le Travail Humain,* 51(2):139–171.

Haley, P. H. and Woods, D. D. (1988). A method and apparatus for dynamic systems control. Patent number 4,770,841, patent date Sept. 13.

Bennett, K. B., Toms, M. L. and Woods, D. D. (1993). Emergent features and configural elements: Designing more effective integral computer based displays of data. *Human Factors*,35:71--98.

Co-Active Architecture

Johnson, M., J.M. Bradshaw, P. J. Feltovich, C. M. Jonker, M. B. van Riemsdijk, and M. Sierhuis. Coactive Design: Designing Support for Interdependence in Joint Activity. Journal of Human-Robot Interaction, Vol. 3, No. 1, (2014), pp. 43-69.

Johnson, M., Bradshaw, J. M., Hoffman, R. R., Feltovich, P. J., & Woods, D. D. (2014). Seven Cardinal Virtues for Human-Machine Teamwork: Examples from the DARPA Robotic Challenge. IEEE Intelligent Systems, November/December 2014 (vol. 29 issue 6), pp. 74-80.

Johnson, M., Bradshaw, J. M., & Feltovich, P. J. (2017). Tomorrow's Human–Machine Design Tools: From Levels of Automation to Interdependencies. *Journal of Cognitive Engineering and Decision Making*.

Schraagen, J.-M., Barnhoorn, J. S., van Schendel, J. & van Vught, W. (2022) Supporting teamwork in hybrid multi-team systems, *Theoretical Issues in Ergonomics Science*, 23:2, 199-220, DOI: 10.1080/1463922X.2021.1936277

Vignati, M., Johnson, M., Bunch, L., Carff, J. Duran, D. (2022) Interdependence design principles in practice . *Frontiers in Physics,* Interdisciplinary Approaches to the Structure and Performance of Interdependent Autonomous Human Machine Teams and Systems (A-HMT-S). 10:969544. 07 December 2022 https://doi.org/10.3389/fphy.2022.969544

Joint Activity

Klein, G., Woods. D.D., Bradshaw, J., Hoffman, R.R., and Feltovich, P.J., (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems,* 19(6), 91-95.

Klein, G., Feltovich, P.J., Bradshaw, J. M. and Woods D. D. (2005). Common ground and coordination in joint activity. In: WB Rouse, and KR Boff, editors. *Organizational Simulation*. Hoboken, NJ, USA: John Wiley & Sons. p. 139–84. doi:10.1002/0471739448.ch6

Maguire, L. M. D. (2020). Managing the hidden costs of coordination. *Communications of the ACM*, 63(4), p. 90-96. https://doi.org/10.1145/3379989

van Diggelen, J., Neerincx, M., Peeters, M. and Schraagen, J.-M. (2019). Developing Effective and Resilient Human-Agent Teamwork Using Team Design Patterns. *IEEE Intelligent Systems*, 34, 2,15-24, March-April, doi: 10.1109/MIS.2018.2886671

Morey, D. A., Marquisee, J. M., Gifford, R. C., Fitzgerald, M. C., & Rayo, M. F. (2020, December). Predicting graceful extensibility of human-machine systems: a new analysis method for evaluating extensibility plots to anticipate distributed system performance. In *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting (Vol. 64, No. 1, pp. 313-318). Sage CA: Los Angeles, CA: SAGE Publications.

Morey, D. A. (2023). Jointness Still Matters: Adding AI Without Designing for Joint Activity Likely Degrades Performance (Doctoral dissertation, The Ohio State University).

Morey, D. A., & Rayo, M. F. (2023). Situated Interpretation and Data: Explainability to Convey Machine Misalignment. *IEEE Transactions on Human-Machine Systems*.

Sample Trajectories of Complexification

at the intersection of people, technology and work.

How sensors extend human perception and attention to new scales

Woods, D. D. and Sarter, N. B. (2010). Capturing the Dynamics of Attention Control From Individual to Distributed Systems. *Theoretical Issues in Ergonomics*, 11(1), 7-28.

Morison, A., Woods, D., Murphy, T. (2015). Human-Robot Interaction as Extending Human Perception to New Scales. In *Cambridge Handbook of Applied Perception Research*; Hoffman, R., Hancock, P., Parasuraman, R., Szalma, J., Scerbo, M., Eds.; Cambridge University Press: Cambridge, UK; pp. 848–868.

Morison, A. and Woods, D. D. (2016). Opening up the Black Box of Sensor Processing Algorithms through New Visualizations. *Informatics*, *3*(3), 16; doi:10.3390/informatics3030016.

Highly Autonomous but Fragile Critical Digital Services at Scale

Allspaw, J. (2016). Human factors and ergonomics practice in web engineering and operations: navigating a critical yet opaque sea of automation. *Human Factors and Ergonomics in Practice*, pp. 313-322, CRC Press (Taylor & Francis) Boca Raton, FL.

Woods, D. D., ed. (2017). STELLA: Report from the SNAFUcatchers Workshop on Coping With Complexity. SNAFU Catchers Consortium, October 4, 2017, Columbus, OH: The Ohio State University, downloaded from stella.report 10/05/2017.

Allspaw, J. and Cook, R. I. (2018). SRE cognitive work. In Blank-Edelman, D. N. (Ed.), *Seeking SRE: Conversations About Running Production Systems at Scale*, O'Reilly, pp. 441-465.

Cook, R. I. (2020). Above the line, below the line. *Communications of the ACM*, 63(3), 43–46 https://doi.org/10.1145/3379510

Woods, D. D. and Allspaw, J. (2020). Revealing the Critical Role of Human Performance in Software. *Communications of the ACM*, 63(5), p. 64-67. DOI:10.1145/3380468.

Cook, R. I. and Long, B. A. (2021). Building and revising adaptive capacity sharing for technical incident response: A case of resilience engineering. *Applied Ergonomics* 90 (2021) 103240

Woods, D. D. (2022). Why do reliable systems fail?. Adobe Research Summit, May 24, 2022. https://www.youtube.com/watch?v=fbwDnpuys7w&t=103s

Allspaw, J. (2022). Critical Digital Services. In H. Laroche, C. Bieder, & J. Villena-L.pez (Eds.), Managing Future Challenges for Safety: Demographic Change, Digitalisation and Complexity in the 2030s (pp. 33–40). Springer International Publishing.https://doi.org/10.1007/978-3-031-07805-7_4

Theoretical Progress on Architecting

Distributed Layered Networks of People and Automata

Robust Yet Fragile Character of Distributed Layered Networks/Human-Machine Systems

Carlson J. M. & Doyle, J. (2000). Highly optimized tolerance: Robustness and design in complex systems. *Physics Review Letters*, 84, 2529-2532.

Cook, R. I. & Rasmussen, J. (2005). Going solid: A model of system dynamics and consequences for patient safety. *Quality & Safety in Health Care*, *14*(2), 130-134.

Woods, D. D. (2006). Essential Characteristics of Resilience for Organizations. In E. Hollnagel, D.D. Woods and N. Leveson, eds., *Resilience Engineering: Concepts and Precepts*. Ashgate, Aldershot, UK, pp. 21-34.

Alderson, D. L. and Doyle, J. C. (2010). Contrasting views of complexity and their implications for network-centric infrastructures. *IEEE SMC—Part A*, 40:839-852.

Woods, D. D. and Branlat, M. (2010). Hollnagel's test: being 'in control 'of highly interdependent multi-layered networked systems. *Cognition, Technology, and Work*, 12(2), 95-101.

Woods, D. D. and Branlat, M. (2011). How Adaptive Systems Fail. In E. Hollnagel, Paries, J., Woods, D.D., and Wreathall, J., Eds., *Resilience Engineering in Practice*. Ashgate, Aldershot, UK, pp. 127-143.

Hoffman, R. R. and Woods, D. D. (2011). Beyond Simon's Slice: Five Fundamental Trade-Offs that Bound the Performance of Human Macrocognitive Work Systems. *IEEE Intelligent Systems*, 26(6), November/December, 67-71.

Eisenberg, D.A., Seager, T.P., and Alderson, D.L. (2019). Rethinking Resilience Analytics, *Risk Analysis*, 39(9): 1870-1884, DOI:10.1111/risa.13328.

Theory: Diversity Enabled Sweet Spots

Anderson, J., Doyle, J. C., Low, S. H. and Matni, N. (2019). System level synthesis. *Annual Reviews in Control*, 47, 364-393.

Nakahira, Y., Liu, Q., Sejnowski, T. J. and Doyle, J. C. (2021). Diversity-enabled sweet spots in layered architectures and speed–accuracy trade-offs in sensorimotor control. *Proceedings of the National Academy of Science PNAS*, 118, 22 https://doi.org/10.1073/pnas.1916367118

Theory of Graceful Extensibility

Woods, D. D. (2015). Four Concepts of Resilience and the Implications for Resilience Engineering. *Reliability Engineering and Systems Safety*, 141, 5-9. published online: 3 APR 2015 doi:10.1016/j.ress.2015.03.018.

Woods, D. D. (2018). The Theory of Graceful Extensibility. *Environment Systems and Decisions*, 38:433–457. https://doi.org/10.1007/s10669-018-9708-3

Woods, D.D. (2019). Essentials of Resilience, Revisited. In M. Ruth and S. G. Reisemann (Eds)., *Handbook on Resilience of Socio-Technical Systems*. Edward Elgar Publishing, pp. 52-65.

Application of Theory of Graceful Extensibility to Human-Autonomy Systems

Woods, D. D. and Balkin, E. A. (2018). A Resiliency Trade Space Study of Detect and Avoid Autonomy on Drones When Communications are Degraded. Report for NASA Ames Research Center June 30, 2018.

Eraslan, E., Yildiz, Y. and Annswamy, A. M. (2020). Shared Control Between Pilots and Autopilots: Illustration of a Cyber-Physical Human System. *IEEE Control Systems Magazine* 40(6):77-97, December 2020. DOI: 10.1109/MCS.2020.3019721

Farjadian, A. B., Thomson, B., Annaswamy, A. M. and Woods, D. D. (2021). Resilient Flight Control: An Architecture for Human Supervision of Automation. *IEEE Transactions on Control Systems Technology*, 29(1), 29-42. DOI: 10.1109/TCST.2019.2959542

View publication stats